

Forecasting High-Dimensional Realized Volatility Matrices Using A Factor Model

Keren Shen [☆]

*Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam, Hong Kong*

Jianfeng Yao

*Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam, Hong Kong*

Wai Keung Li

*Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam, Hong Kong*

Abstract

Modeling and forecasting covariance matrices of asset returns play a crucial role in finance. The availability of high frequency intraday data enables the modeling of the realized covariance matrix directly. However, most models in the literature suffer from the curse of dimensionality. To solve the problem, we propose a factor model with a diagonal CAW model for the factor realized covariance matrices. Asymptotic theory is derived for the estimated parameters. In an extensive empirical analysis, we find that the number of parameters can be reduced significantly. Furthermore, the proposed model maintains a comparable performance with a benchmark vector autoregressive model.

Keywords: High-dimension, High-frequency, Realized covariance matrices, Factor model, Wishart distribution

JEL: C31, C38, C55

[☆]Corresponding Author:

Email address: rayshenkr@hku.hk (Keren Shen [☆])

1. Introduction

Modeling and forecasting covariances or volatility matrices of asset returns play a crucial role in many financial fields, such as portfolio allocation (Markowitz, 1952) and asset pricing (Bollerslev et al., 1988). With the availability of intraday financial data nowadays, it becomes possible to estimate volatilities and co-volatilities of asset returns using high-frequency data directly, which leads to the so-called *realized covariance matrix* (Andersen et al., 2003, Barndorff-Nielsen and Shephard, 2004 and Barndorff-Nielsen et al., 2011). Two major problems arise in the estimation of realized covariance matrices. Firstly, transactions for different assets are typically asynchronous so that the high-frequency prices of different assets do not change simultaneously. Secondly, it is widely believed that the observed high-frequency prices are accompanied by *microstructure noise* so that the observed prices should be thought as a noisy version of the true underlying price process. Researchers have proposed several ways to tackle these problems, for example, the overlap intervals and previous tick method by Hayashi and Yoshida (2005) and Zhang (2011), respectively. Moreover, Bannouh et al. (2012) use a refresh time scheme and Christensen et al. (2010) propose the pre-averaging approach.

Once constructed, realized covariance matrices are analyzed using multivariate models. There are two major issues to be resolved in modeling realized covariance matrices. The first issue is that the model should guarantee the positive definiteness of fitted covariance matrices. A natural choice in this aspect is the family of matrix-valued Wishart distributions which automatically generates random positive definite matrices without imposing additional constraints. Several models related to the Wishart distribution have been put forward. Gourieoux et al. (2009) propose the Wishart Autoregressive (WAR) model where the realized covariance matrix has a conditional distribution which is noncentral Wishart with a non-centrality parameter depending on lagged covariances and a fixed scaling matrix. Later, Golosnoy et al. (2012) propose the Conditional Autoregressive Wishart (CAW) model under which the conditional distribution is central Wishart with time dependent scaling matrices. Moreover, Ng et al. (2014) generalize the above two models to construct Generalized Conditional Autoregressive Wishart (GCAW) model. There are other models involving the Wishart distribution, for instance, see Jin and Maheu (2009).

The second issue in modeling realized covariance matrices is the high-dimensionality. Indeed, covariance matrices have $d(d + 1)/2$ entries for d assets; consequently, the number of parameters in the model for realized covariance matrices grows quickly with d . For example, for an unrestricted CAW(2,2) model with $d = 10$ assets, as many as 456 parameters are needed so that it is quite challenging to fit such a model in practice. This is probably the major reason why all the empirical studies we find in the literature on model fitting for realized covariance matrices are all limited to a *small number*, say 3 to 5 assets. Another problem inherent in realized covariance matrices is that they deviate from their population counterpart, the so-called *integrated covariance matrix*, when the number of assets is large compared to the sample size (Johnstone and Lu, 2009; and Wang and Zou, 2010). It is therefore important to build statistical models for realized covariance matrices with a large dimension, say several tens.

Improved estimators of realized covariance matrices are proposed in Wang and Zou (2010) with a so-called averaging realized volatility matrix (ARVM) estimator. Tao et al. (2011) propose the threshold averaging realized volatility matrix (TARVM) estimator which is of two-scale and uses the previous-tick method and the threshold technique in constructing realized volatility matrices. Then, inspired by Zhang (2006) and Fan and Wang (2007), Tao et al. (2013) propose the threshold multi-scale realized volatility matrix (TMSRVM) estimator. The TARVM and TMSRVM estimators are shown to be consistent for the integrated covariance matrix when the dimension of the realized covariance matrix, the sample size of intraday points and the length of sampling days go to infinity. In addition, the TMSRVM estimator proves to have the optimal convergence rate under the existence of the microstructure noise.

As an effort to control the parametric dimension of the models, Tao et al. (2011) propose to first identify a small number of factors for the realized covariance matrices and to fit a vector autoregressive (VAR) model to the vectorized factor covariance matrices. They show that such a factor model significantly reduces the number of parameters needed to fit the realized covariance matrices. However, the VAR specification is not able to ensure the positive definiteness of the predicted factor covariance matrices. In addition (see below), such a VAR fit still needs $O(r^4)$ number of parameters with r factors.

In this article, we adopt the factor model approach introduced in Tao et al. (2011) for re-

alized covariance matrices. In order to overcome the aforementioned weakness of their VAR fit for the extracted factors, we propose a diagonal CAW model which has several advantages. Firstly, the proposed CAW model is able to guarantee automatically the positive definiteness of the covariance matrices generated from the model without imposing additional constraints. Secondly, as will be shown by extensive data analysis reported in this paper, our model has excellent empirical performance in terms of the reduction of number of parameters compared to the VAR approach proposed in [Tao et al. \(2011\)](#): indeed we obtain comparable forecasting performance with much less parameters.

In a related work, [Asai and McAleer \(2014\)](#) also use a combination of factor extraction and CAW modeling and report some empirical studies with 7 assets which is still considered to be small dimension. In this paper, we focus on a larger collection of assets where empirical studies are carried out for 30 assets. A further difference in this paper is that we also propose a thorough theoretical analysis of both the factor modeling and the CAW estimation.

The rest of the paper is organized as follows. Section 2 introduces the model setup and our approach based on a factor model and a diagonal CAW model for the extracted factors. In Section 3, the asymptotic theory is established. In Sections 4 and 5, we report the middle and large scale data analysis on asset prices, respectively. Conclusions are presented in Section 6. Proofs of the asymptotic theory are provided in the Appendix.

2. Methodology

2.1. Model setup

Suppose there are d assets and their log price process $\mathbf{X}(t) = \{X_1(t), \dots, X_d(t)\}'$ follows a continuous diffusion model:

$$d\mathbf{X}(t) = \mu_t dt + \sigma_t d\mathbf{W}_t, \quad t \in [0, T], \quad (1)$$

where μ_t is a drift in \mathbb{R}^d , \mathbf{W}_t a standard d -dimensional Brownian motion and σ_t a $d \times d$ matrix. The *integrated volatility matrix* for the t -th day is defined as

$$\Sigma_x(t) = \int_{t-1}^t \sigma_s \sigma_s' ds, \quad t = 1, \dots, T. \quad (2)$$

However, it is commonly admitted that the microstructure noise is inherent in the high-frequency price process so that we are not able to observe directly $X_i(t)$, but $Y_i(t_{i\ell})$, a noisy version of $X_i(\cdot)$ at times $t_{i\ell}$, $\ell = 1, \dots, n_i$, $i = 1, \dots, d$. Here, n_i is the total trading times and $t_{i\ell}$ is the ℓ -th trading times of asset i during a giving trading day t . The observations $Y_i(t_{i\ell})$ is allowed to be non-synchronized, i.e. $t_{i\ell} \neq t_{j\ell}$ for any $i \neq j$. In this paper, we assume that

$$Y_i(t_{i\ell}) = X_i(t_{i\ell}) + \epsilon_i(t_{i\ell}) \quad (3)$$

where $\epsilon_i(t_{i\ell})$ are i.i.d. microstructure noise with mean zero and variance η_i , and $\epsilon_i(\cdot)$ and $X_i(\cdot)$ are independent with each other.

2.2. Realized covariance matrix estimator

Several issues arise for the estimation of $\Sigma_x(t)$: 1. asynchronous observations of different assets; 2. microstructure noise; 3. the number of assets can be larger than the sample size. In this paper, we adopt the threshold Multi-Scale Realized Volatility Matrix estimator (threshold MSRVM) proposed by [Tao et al. \(2013\)](#), denoted by $\hat{\Sigma}_x(t)$, $t = 1, \dots, T$. The threshold MSRVM estimator has many attractive properties, for instance, it is consistent for the high-dimensional integrated co-volatility matrix with the optimal convergence rate. Briefly, the idea of the threshold MSRVM estimator is the following: the previous-tick method is used to construct the raw realized covariance matrices. Then, a multi-scale estimator is evaluated which is actually a kind of average of those raw estimators. In addition, the multi-scale estimator is regularized using a thresholding method, that is, matrix entries under a threshold are set to be zero.

2.3. Matrix factor model

We adopt the factor model proposed by [Tao et al. \(2011\)](#) to reduce the large dimension of $\Sigma_x(t)$:

$$\Sigma_x(t) = \mathbf{A}\Sigma_f(t)\mathbf{A}' + \Sigma_0, \quad (4)$$

for $t = 1, \dots, T$, where $\Sigma_f(t)$ are $r \times r$ positive definite factor covariance matrices, Σ_0 is a $d \times d$ positive definite constant matrix and \mathbf{A} is a $d \times r$ factor loading matrix normalized by the constraint $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$. As in a standard factor model, only the left-hand side of Equation (4) is observed. The unknown quantities are estimated using the method in [Tao et al. \(2011\)](#). Let

$$\bar{\Sigma}_x = \frac{1}{T} \sum_{t=1}^T \Sigma_x(t), \quad \bar{S}_x = \frac{1}{T} \sum_{t=1}^T \{\Sigma_x(t) - \bar{\Sigma}_x\}^2, \quad (5)$$

and

$$\tilde{\Sigma}_x = \frac{1}{T} \sum_{t=1}^T \hat{\Sigma}_x(t), \quad \tilde{S}_x = \frac{1}{T} \sum_{t=1}^T \{\hat{\Sigma}_x(t) - \tilde{\Sigma}_x\}^2. \quad (6)$$

Next, the estimator $\hat{\mathbf{A}}$ is obtained using the r orthonormal eigenvectors of \tilde{S}_x , corresponding to its r largest eigenvalues, as its columns. Finally, the estimated factor covariance matrices are

$$\hat{\Sigma}_f(t) = \hat{\mathbf{A}}'\hat{\Sigma}_x(t)\hat{\mathbf{A}} \quad (7)$$

for $t = 1, \dots, T$; and Σ_0 is estimated by

$$\hat{\Sigma}_0 = \tilde{\Sigma}_x - \hat{\mathbf{A}}\hat{\mathbf{A}}'\tilde{\Sigma}_x\hat{\mathbf{A}}\hat{\mathbf{A}}'. \quad (8)$$

2.4. CAW modeling for factor covariance matrix

With the estimated factor covariance matrices $\{\hat{\Sigma}_f(t)\}$ calculated by (7) and (8), we construct a dynamic structure by fitting a diagonal CAW model to $\tilde{\Sigma}_f(t)$, where $\tilde{\Sigma}_f(t) := \Sigma_f(t) + \mathbf{A}'\Sigma_0\mathbf{A}$. Here, $\tilde{\Sigma}_f(t)$ is modeled rather than $\Sigma_f(t)$ since $\hat{\Sigma}_f(t)$ is in fact a consistent estimator of the former, and it is impossible to construct a consistent estimator for the latter, to our knowledge.

The model is defined as follows. Let $\mathcal{F}_{t-1} = \sigma(\tilde{\Sigma}_f(s), s < t)$ be the past history of the

process at time t . Conditional on \mathcal{F}_{t-1} , $\tilde{\mathbf{S}}_f(t)$ follows a central Wishart distribution

$$\tilde{\mathbf{S}}_f(t)|\mathcal{F}_{t-1} \sim \mathcal{W}_n(\nu, \mathbf{S}_f(t)/\nu), \quad (9)$$

with ν the degrees of freedom and the scaling matrix $\mathbf{S}_f(t)$. Moreover, the scaling matrix $\mathbf{S}_f(t)$ follows a linear recursion of order (p, q)

$$\mathbf{S}_f(t) = CC' + \sum_{i=1}^p B_i \mathbf{S}_f(t-i) B_i' + \sum_{j=1}^q A_j \tilde{\mathbf{S}}_f(t-j) A_j', \quad (10)$$

where A_j , B_i and C are all $r \times r$ matrices of coefficients.

In summary, the CAW process depends on the parameters $\{\nu, C, (B_i)_{1 \leq i \leq p}, (A_j)_{1 \leq j \leq q}\}$ without additional constraints, so that the total number of parameters is equal to $(p+q)r^2 + \frac{r(r+1)}{2} + 1 = O(r^2)$ which still grows quickly with the number of factors r and the order p and q . Since the main aim of the paper is to propose a practically feasible model for a large number of assets while retaining efficiency, we will restrict ourselves to *diagonal* coefficient matrices C , $(B_i)_{1 \leq i \leq p}$ and $(A_j)_{1 \leq j \leq q}$. Therefore, the number of parameters becomes $(p+q+1)r + 1 = O(r)$. Notice that this setup is also supported by the fact that in the literature ([McCurdy and Stengos, 1992](#), [Engle and Kroner, 1995](#)), researchers tend to use diagonal volatility models to avoid overparameterization and argue that the variances and the covariances rely more on its own past than the history of other variances or covariances. In the empirical study developed below, we find that the diagonal models achieve a comparable performance with unrestricted ones, while being much more parsimonious and requiring far less computing time. Notice however, the asymptotic theory developed below is also valid for unrestricted matrices A_j 's, B_i 's and C .

The estimation of the parameters $\theta = (\nu, \text{diag}(C)', \text{diag}(B_i)'_{1 \leq i \leq p}, \text{diag}(A_j)'_{1 \leq j \leq q})'$ of the diagonal CAW(p, q) model is carried out by maximizing the log-likelihood function using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization procedure. Positivity of the diagonal elements C_{kk} , $A_{11,j}$ and $B_{11,i}$ are enforced, where $1 \leq k \leq r$. The log-likelihood function is

$$\mathcal{L}(\theta) = \frac{1}{T} \sum_{t=1}^T \left\{ -\frac{\nu r}{2} \ln(2) - \frac{r(r-1)}{4} \ln(\pi) - \sum_{i=1}^r \ln \Gamma\left(\frac{\nu+1-i}{2}\right) - \frac{\nu}{2} \ln |\mathbf{S}_f(t)| \right\}$$

$$+(\frac{\nu-r-1}{2})\ln|\tilde{\Sigma}_f(t)| - \frac{1}{2}\text{tr}(\nu\mathbf{S}_f(t)^{-1}\tilde{\Sigma}_f(t))\} \quad (11)$$

In practice, initial values for $\mathbf{S}_f(t)$ are needed to run the maximization of the log-likelihood function. For example, if the order $(p, q) = (2, 2)$ is used, then the initial values $\mathbf{S}_f(1)$ and $\mathbf{S}_f(2)$ are needed. In the empirical analysis of this paper, we take $\mathbf{S}_f(1) = \hat{\Sigma}_f(1)$ and $\mathbf{S}_f(2) = \hat{\Sigma}_f(2)$ as $\mathbf{S}_f(t)$ is the conditional expectation of $\tilde{\Sigma}_f(t)$, for any t .

3. Asymptotic theory

Given a d -dimensional vector $\mathbf{x} = (x_1, \dots, x_d)'$ and a $d \times d$ matrix $\mathbf{U} = (U_{ij})$, define vector and matrix norms as

$$\|\mathbf{x}\|_2 = (\sum_{i=1}^d |x_i|^2)^{1/2}, \quad \|\mathbf{U}\|_2 = \sup\{\|\mathbf{U}\mathbf{x}\|_2, \|\mathbf{x}\|_2 = 1\} \quad (12)$$

In fact, $\|\mathbf{U}\|_2$ is the spectral norm, equal to the square root of the largest eigenvalue of $\mathbf{U}'\mathbf{U}$. In addition, define the Frobenius norm of the $d \times d$ matrix $\mathbf{U} = (U_{ij})$ as

$$\|\mathbf{U}\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^d |U_{ij}|^2} \quad (13)$$

The asymptotic theory below uses the following assumptions.

(A1) All row vectors of \mathbf{A}' and Σ_0 in the factor model (4) satisfy the sparsity condition (13) below. We say that a d -dimensional vector $\mathbf{x} = (x_1, \dots, x_d)'$ is sparse if

$$\sum_{j=1}^d |x_j|^\delta \leq C\pi(d), \quad (14)$$

where $0 \leq \delta < 1$, $\pi(d)$ is a deterministic function of d that grows slowly in d , such as $\pi(d) = 1$ or $\ln(d)$, and C is a positive constant.

(A2) The factor model (4) has r fixed factors, with $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$, and matrices Σ_0 and Σ_f satisfy

$$\begin{aligned} \|\Sigma_0\|_2 &< \infty \\ \max_{1 \leq t \leq T} |\Sigma_{f,jj}(t)| &= O_p(B(T)), \quad j = 1, \dots, r. \end{aligned} \quad (15)$$

where $1 \leq B(T) = o(T)$.

(A3) $\max_{1 \leq t \leq T} \|\hat{\Sigma}_x(t) - \Sigma_x(t)\|_2 = O_p(A(d, T, n))$, for some rate function $A(d, T, n)$ such that $A(d, T, n)B^5(T) = o(1)$ with $B(T)$ defined in (A2).

(A4) A_j 's and B_i 's are such that the CAW model (9) and (10) are stationary and ergodic.

(A5) The parameter set Θ for all parameters A_j 's, B_i 's, C and ν is compact for the CAW model (9) and (10).

(A6) The Hessian matrix $\partial^2 \mathcal{L}(\theta) / \partial \theta_i \partial \theta_j$ converges to some deterministic matrix function $D(\theta)$ as T goes to infinity which is of full rank for all $\theta \in \Theta$.

The first two conditions are from [Tao et al. \(2011\)](#) which are used to prove the consistency of $\hat{\Sigma}_f(t)$. In this paper, as we are using the threshold MSRVM estimator, we take $A(d, T, n) = \pi(d)[e_n(d^2 T)^{1/\beta}]^{1-\delta} \ln T$ and $B(T) = \ln T$, where $e_n \sim n^{-1/4}$, consistent with [Tao et al. \(2013\)](#). In addition, the constant β can be taken large enough (under reasonable moment conditions) so that $A(d, T, n)B^5(T)$ will go to 0 as n, d , and T go to infinity. Consequently, we assume here these two values shall converge to 0 in some sense.

Theorem 1. *Suppose the models (1), (3) and (4) satisfy Conditions (A1)-(A3). Denote the ordered eigenvalues of $\bar{\mathbf{S}}_x$ by $\lambda_1 \geq \dots \geq \lambda_d$. Let $\mathbf{a}_1, \dots, \mathbf{a}_r$ be the eigenvectors of $\bar{\mathbf{S}}_x$ corresponding to the r largest eigenvalues $\lambda_1, \dots, \lambda_r$. Also set $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_r$ be the r largest eigenvalues of $\hat{\bar{\mathbf{S}}}_x$ and $\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_r$ the corresponding eigenvectors. Let $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_r)$ and $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_r)$. As n, d, T go to infinity, we have*

$$\begin{aligned} \mathbf{A}'\hat{\mathbf{A}} - \mathbf{I}_r &= O_p(A(d, T, n)B(T)), \\ \hat{\Sigma}_f(t) - \Sigma_f(t) - \mathbf{A}'\Sigma_0\mathbf{A} &= O_p(A^{1/2}(d, T, n)B^{3/2}(T)). \end{aligned} \tag{16}$$

Theorem 2. *Suppose that $\hat{\theta}$ is the maximized log-likelihood estimator of θ based on the data $\hat{\Sigma}_f(t)$ from the CAW model and $\tilde{\theta}$ is the maximized log-likelihood estimator based on the true*

data $\tilde{\Sigma}_f(t)$ from the same CAW model. Then under Conditions (A1)-(A5),

$$\hat{\theta} - \tilde{\theta} = O_p(A^{1/2}(d, T, n)B^{5/2}(T)) \quad (17)$$

4. Data analysis 1

We apply the proposed methodology to two datasets. In this section, we focus on comparing the VAR and CAW models, using 5-minutes intraday data of 30 stocks traded in the US stock market over a period of 103 days where the data are obtained directly from Bloomberg.

4.1. Data description

We use 30 stocks traded at the New York Stock Exchange, which consist of 27 components of Dow Jones Index : 3M (MMM), American Express (AXP), AT&T (T), Boeing (BA), Caterpillar (CAT), Chevron (CVX), Coca-Cola (KO), Dupont (DD), ExxonMobil (XOM), General Electric (GE), Goldman Sachs (GS), The Home Depot (HD), IBM (IBM), Johnson & Johnson (JNJ), JPMorgan Chase (JPM), McDonald's (MCD), Merck (MRK), Nike (NKE), Pfizer (PFE), Procter & Gamble (PG), Travelers (TRV), UnitedHealth Group (UNH), United Technologies (UTX), Verizon (VZ), Visa (V), Wal-Mart (WMT), Walt Disney (DIS) and three former components of Dow Jones Index: Honeywell (HON), Citigroup (C) and American International Group (AIG). The daily realized covariance matrix is computed as $\hat{\Sigma}_x(t) = \sum_{j=1}^M y_{t,j} y'_{t,j}$, where $y_{t,j}$ is the vector of log-returns for the 30 stocks computed for the j th 5-minute interval of trading day t except the first and last half hour. The sample period starts at May 3, 2013 and ends on September 30, 2013, totally 103 trading days. (Here, we exclude July 4 due to the incompleteness of data). This generates a series of 103 matrices of $\hat{\Sigma}_x(t)$, which are 30 by 30.

Descriptive statistics of the realized variances and covariances are provided in Table 1. We only show some entries due to limited space. The following properties are found:

1. Among 30 realized variances and 435 covariances, only 3 realized covariances are skewed to the left rather than to the right.

2. All realized variances and covariances have bigger kurtosis than that of the normal distribution except for 2 realized covariances.

For the next two subsections, we show the estimated results for both the diagonal CAW and VAR models when the first $k = 98$ days are treated as data points.

4.2. Model fitting

The eigenvalues of the sample variance matrix $\bar{\mathbf{S}}_x$ are evaluated, and are shown in Figure 1. The plots show that the three largest eigenvalues are much larger than the others, which indicates that the factor number $r = 3$ is appropriate.

Let $\hat{\mathbf{A}}$ be the eigenvector of $\bar{\mathbf{S}}_x$ corresponding to the three largest eigenvalues. We then calculate the factor volatility matrices $\hat{\Sigma}_f(t)$, which are 3 by 3. Figure 2 shows time series plots of the variances and covariances of the factor covariance matrices.

Then, we fit the diagonal CAW model to the matrix series. Different orders are used to make comparison, namely $(p, q) = (0, 1)$, $(p, q) = (1, 1)$, $(p, q) = (1, 2)$, $(p, q) = (2, 1)$ and $(p, q) = (2, 2)$.

Remark 1. *The initial value for each numerical optimization is chosen randomly and we repeat 160 times to choose the one with the largest log-likelihood value.*

4.3. VAR model

For comparison purpose, we also fit the VAR model, advocated in Tao et al. (2011), to the vectorized factor covariance matrix $\hat{\Sigma}_f(t)$, which is a vector with 6 entries. Using the package "vars" in R, we select the VAR(1) model as all the model selection criteria, namely Akaike information criterion (AIC), Hannan-Quinn (HQ), Schwarz criterion (SC) and final prediction error (FPE), choose the order 1, as shown in Table 2.

The model is

$$\text{vech}\{\tilde{\Sigma}_f(t)\} = \mathbf{A}_0 + \mathbf{A}_1 \text{vech}\{\tilde{\Sigma}_f(t-1)\} + \mathbf{e}(t)$$

where \mathbf{A}_0 is a 6-dimensional vector, \mathbf{A}_1 is a 6×6 square matrix, and $\mathbf{e}(t)$ is a 6-dimensional vector white noise process with zero mean and finite fourth moments. Both \mathbf{A}_0 and \mathbf{A}_1 are estimated by the least squares method.

The fitted VAR(1) model has the coefficients:

$$\hat{\mathbf{A}}_0 = 10^{-5} \begin{pmatrix} 54.49 \\ -1.482 \\ 4.937 \\ 7.632 \\ 0.164 \\ 10.27 \end{pmatrix}, \quad \hat{\mathbf{A}}_1 = 10^{-2} \begin{pmatrix} 55.8 & -31.2 & 19.6 & -37.4 & -199.1 & -8.0 \\ -4.1 & 20.6 & 3.2 & 7.7 & 45.4 & -2.0 \\ 6.8 & 15.3 & 72.8 & -6.5 & -230.9 & -0.1 \\ 0.5 & -26.5 & -84.6 & 4.4 & 283.8 & -6.0 \\ -0.6 & 2.5 & 2.2 & 1.0 & -3.7 & -0.3 \\ 10.1 & 43.5 & -17.4 & 11.1 & -207.5 & 13.4 \end{pmatrix}$$

The fitted models are shown in Figure 4 which indicates that the VAR(1) provides adequate fit to the data. For each of the 6 time series, we show the in-sample fit in the upper panel, where the solid line stands for the real data and the dashed line is the fitted series. The residuals are shown in the middle panel while the ACF and PACF of residuals are displayed in the lower panel. The ACF and PACF plots show that the residuals are time-uncorrelated.

4.4. Performance comparison in out-of-sample forecasting

We compare the out-of-sample one-day-ahead forecast performance of the diagonal CAW and VAR models. The one-day-ahead realized covariance is calculated by: 1. Obtain the one-day-ahead factor covariance matrix by conditional expectation; 2. Plug the forecast factor covariance matrix into the factor model (4) to get the predicted realized covariance matrix.

The predictive accuracy is measured with both the Frobenius norm and the spectral norm. We take the first k days as data and forecast the next day, where $k = 80, \dots, 98$. Every model is re-estimated and the new forecasts are generated based upon the new parameter estimates. Then, we take the average of errors during 19 periods to do the comparison. Moreover, the errors of the inverse of matrices are also compared.

Table 3 contains the results of the prediction accuracy of the two models using different norms. The main findings are as follows.

1. The CAW models with order $(p, q) = (1, 1)$ performs the best as it has the smallest error under both Frobenius and spectral norms. In general, All CAW models have similar performance except the ones with order $(p, q) = (0, 1)$.
2. The CAW models have similar performance with that of the VAR(1) model except the one with order $(p, q) = (0, 1)$.
3. However, the diagonal CAW model needs far less parameters than the VAR model does. Here, the best CAW model with order $(p, q) = (1, 1)$ only needs 10 parameters, just a quarter of the number of parameters that the VAR(1) model needs.
4. As d goes up, we can imagine that r will become larger; in this situation, the parameters that we need for the diagonal CAW model will be even far less than that we need for the VAR model.

Remark 2.

1. *We do predictions for 2 to 5 days ahead in addition to the one-day forecast. We find that the performance of predictions for longer horizons is similar with that for the one day, in general.*
2. *One problem for the VAR model is that it cannot assure that the predicted factor covariance matrix is positive definite. We have checked that in this data analysis, the predicted factor covariance matrices are actually all positive definite. This may be due to the reason that the estimated factor variances are larger than the covariances (in absolute value).*

5. Data analysis 2

5.1. Data description

We use the same 30 stocks as in the previous section. The raw tick-by-tick trading data are downloaded from TAQ database of Wharton Research Data Service. The data period starts at January 3, 2012 and ends on December 31, 2012, with totally 250 trading days.

We firstly conduct data cleaning with the procedures introduced in [Brownlees and Gallo \(2006\)](#) and [Barndorff-Nielsen et al. \(2009\)](#). The steps are the following:

1. Delete entries with a time stamp outside 9:30 am - 4:00 pm when the exchange is open.
2. Delete entries with a time stamp inside 9:30 - 10:00 am or 3:30 - 4:00 pm to eliminate the open and end effect of price fluctuation.
3. Delete entries with the transaction price equal to zero.
4. If multiple transactions have the same time stamp, use the median price.
5. Delete entries with prices which are outliers. Let $\{p_i\}_{i=1}^N$ be an ordered tick-by-tick price series. We treat the i -th price as an outlier if $|p_i - \bar{p}_i(k)| > 3s_i(k)$, where $\bar{p}_i(k)$ and $s_i(k)$ denote the sample mean and sample standard deviation of a neighborhood of k observations around i , respectively. For the beginning prices which may not have enough left hand side neighbors, we get $k - i$ neighbors from $i + 1$ to $k + 1$. Similar procedures are taken for the ending prices.

Then, we construct the threshold MSRVM estimator based on the cleaned tick-by-tick data following the steps in [Tao et al. \(2013\)](#). We set the threshold to be 5% of the largest of the absolute value of entries in the matrix.

Descriptive statistics of selected realized variances and covariances are provided in [Table 4](#). In addition, two plots for realized variances and two plots for realized covariances are shown in [Figure 5](#). We find the following properties:

1. All 30 realized variances and 435 covariances are skewed to the right, with mean skewness 1.39.
2. All realized variances and covariances have bigger kurtosis than that of the normal distribution, with mean kurtosis 5.92 showing fat tails.
3. The realized variances and covariances have significant fluctuations during the year, indicated by the graphs.

For the next two subsections, we show the estimated results for both the diagonal CAW and VAR models when the first $k = 220$ days are treated as data points.

5.2. Model fitting

The eigenvalues of the sample variance matrix $\tilde{\mathbf{S}}_x$ are evaluated, and are shown in Figure 5. We choose four factors for the model, as there is a discernable drop between the fourth and the fifth eigenvalues, though the second to fourth eigenvalues are much less than the biggest one.

Let $\hat{\mathbf{A}}$ be the eigenvectors of $\tilde{\mathbf{S}}_x$ corresponding to the four largest eigenvalues. We calculate the factor volatility matrices $\hat{\Sigma}_f(t)$, which are 4 by 4. Then, we fit the diagonal CAW model to the matrix series. Different orders are used to make comparison, namely $(p, q) = (0, 1)$, $(p, q) = (1, 1)$, $(p, q) = (1, 2)$, $(p, q) = (2, 1)$ and $(p, q) = (2, 2)$. For every estimation, we randomly choose 60 initial values for each optimization of the log-likelihood and choose the one with the largest log-likelihood value to give the estimated parameters.

5.3. VAR model

For comparison purpose, we also fit the VAR model to the vectorized factor covariance matrix $\hat{\Sigma}_f(t)$, which is a vector with 10 entries. Again, using the package "vars" in R, we select the VAR(1) model as all the model selection criteria AIC, HQ, SC and FPE choose the order 1, as shown in Table 5.

The model is

$$\text{vech}\{\tilde{\Sigma}_f(t)\} = \mathbf{A}_0 + \mathbf{A}_1 \text{vech}\{\tilde{\Sigma}_f(t-1)\} + \mathbf{e}(t)$$

where \mathbf{A}_0 is a 10-dimensional vector, \mathbf{A}_1 is a 10×10 square matrix, and $\mathbf{e}(t)$ is a 10-dimensional vector white noise process with zero mean and finite fourth moments. Both \mathbf{A}_0 and \mathbf{A}_1 are estimated by the least squares method.

We denote the estimator of \mathbf{A}_1 by $\hat{\mathbf{A}}_1$. We find that $|\hat{\mathbf{A}}_1| = -1.03 \times 10^{-7}$ and the biggest absolute eigenvalues is $0.4906 < 1$ which ensures the stationarity of the VAR model.

In addition, the sparsity of $\hat{\mathbf{A}}_1$ is checked. We compute the values $\frac{1}{10^2} \sum_{i,j} |a_{i,j}|^m$ for different $0 \leq m < 1$, where $\hat{\mathbf{A}}_1 = \{a_{i,j}\}_{1 \leq i,j \leq 10}$. The following is the result. We can see from Table 6 that

the average absolute value of the entries of $\hat{\mathbf{A}}_1$ is small (less than 1) so that we can consider $\hat{\mathbf{A}}_1$ as almost sparse.

5.4. Performance comparison in out-of-sample forecasting

We compare the out-of-sample one-day-ahead forecast performance of the diagonal CAW and VAR models. The prediction of the one-day-ahead realized covariance is calculated by: 1. Predict the one-day-ahead factor covariance matrix by conditional expectation; 2. Plug the forecast factor covariance matrix into the factor model (4) to get the predicted realized covariance matrix.

The predictive accuracy is measured with both the Frobenius norm and the spectral norm. We take the first k days as data and forecast the next day, where $k = 220, \dots, 240$. Every model is re-estimated and the new forecasts are generated based upon the new parameter estimates. Then, we take the average of errors during 21 periods to do the comparison. Moreover, the errors of the inverse of matrices are also compared.

Table 7 contains the results of the prediction error of two models using different norms. The main findings are as follows.

1. The diagonal CAW models with order $(p, q) = (1, 1)$ performs the best among the CAW models as it has the smallest error under both Frobenius and spectral norms. In general, all CAW models have similar performance except the ones with order $(p, q) = (0, 1)$. The result also indicates the possibility of over-parameterization with orders larger than $(1, 1)$.
2. The diagonal CAW models have slightly worse, but comparable performance with that of the VAR(1) model except the one with order $(p, q) = (0, 1)$.
3. The diagonal CAW model needs far less parameters than the VAR model does. Here, the best CAW model with order $(p, q) = (1, 1)$ only needs 13 parameters, nearly a tenth of the number of parameters that VAR(1) model needs.
4. We do predictions for 2 to 5 days ahead in addition to one-day forecast. We find that

the performance of predictions of longer horizons is similar with that for the one day, in general.

6. Conclusions

In the literature, most models dealing with the realized covariance matrix focus on small number of assets, which become infeasible when the dimension is large. In order to solve the problem, we propose a factor model with diagonal CAW model fitted to the factor covariance matrix. Our model performs comparably with the VAR model while requiring far less parameters. For example, in the second data analysis, the CAW model with order $(p, q) = (1, 1)$ performs similarly with the VAR model, measured both in the Frobenius norm and the spectral norm, but only needs nearly one tenth of the number of parameters of the latter. In addition, the model ensures the positive definiteness for the predicted covariance matrices. Diagnostic tests for the proposed model is worth considering in a future study.

Table 1: Descriptive Statistics for Selected Realized Variances and Covariances

We report the descriptive statistics of the realized variances and covariances of the first dataset, namely mean, maximum, minimum, standard deviation, skewness and kurtosis. We only show some entries due to limited space.

Stock	Mean $\times 10^{-5}$	Maximum $\times 10^{-4}$	Minimum $\times 10^{-5}$	SD $\times 10^{-5}$	Skewness	Kurtosis
Realized Variance						
MMM	3.25	1.27	0.63	1.96	1.85	4.84
AXP	6.90	6.05	2.03	7.02	4.74	31.0
T	4.85	1.36	1.29	2.60	1.15	0.85
BA	8.60	28.5	1.31	27.8	9.61	92.8
CAT	6.39	3.70	1.99	4.82	4.43	23.8
Realized Covariance						
MMM-AXP	2.17	1.23	-0.25	2.03	2.13	6.27
MMM-T	1.40	0.77	-0.37	1.48	1.86	3.80
MMM-BA	1.94	0.95	-2.33	1.85	1.36	2.67
MMM-CAT	2.11	1.08	-0.03	1.72	2.23	7.01
AXP-T	1.96	0.96	-0.66	1.95	1.70	3.15
AXP-BA	2.18	1.17	-7.27	2.39	0.45	3.36
AXP-CAT	2.24	1.14	-0.69	2.11	1.80	4.37
T-BA	1.38	0.86	-1.47	1.59	1.95	5.33
T-CAT	1.41	0.84	-1.13	1.61	1.80	3.95
BA-CAT	2.12	0.98	-0.32	1.76	1.65	3.98

Table 2: The Selection of the Order of VAR Model

We fit the VAR model to the vectorized factor covariance matrix $\hat{\Sigma}_f(t)$, which is a vector with 6 entries. Using the package "vars" in R, we select the VAR(1) model by Akaike information criterion (AIC), Hannan Quinn (HQ), Schwarz criterion (SC) and final prediction error (FPE).

Order	1	2	3	4	5
AIC(n) *10 ²	-1.107	-1.102	-1.098	-1.094	-1.090
HQ(n) *10 ²	-1.102	-1.093	-1.084	-1.077	-1.068
SC(n) *10 ²	-1.094	-1.079	-1.065	-1.051	-1.036
FPE(n) *10 ⁻⁴⁸	0.830	1.313	2.181	3.268	5.772

Table 3: Forecast errors for CAW and VAR models using different norms

We report the results of the prediction accuracy of the two models using different norms. In addition, the prediction accuracy of the inverse of the matrices is shown as well. Here, FN is for the Frobenius norm and SN for the spectral norm.

CAW					
Order	Number of Parameters	FN $\ast 10^{-4}$	SN $\ast 10^{-4}$	FN for Inverse $\ast 10^5$	SN for Inverse $\ast 10^5$
$p = 0, q = 1$	7	6.21	5.66	6.30	3.84
$p = 1, q = 1$	10	4.77	3.95	6.31	3.84
$p = 1, q = 2$	13	4.92	4.01	6.31	3.84
$p = 2, q = 1$	13	4.96	4.01	6.31	3.84
$p = 2, q = 2$	16	4.98	4.03	6.31	3.84
VAR					
	Number of Parameters	FN $\ast 10^{-4}$	SN $\ast 10^{-4}$	FN for Inverse $\ast 10^5$	SN for Inverse $\ast 10^5$
VAR(1)	42	4.81	3.99	6.31	3.84

Table 4: Descriptive Statistics for Some Selected Realized Variances and Covariances

We report the descriptive statistics of the realized variances and covariances of the second dataset, namely mean, maximum, minimum, standard deviation, skewness and kurtosis. We only show some entries due to limited space.

Stock	Mean $\ast 10^{-5}$	Maximum $\ast 10^{-4}$	Minimum $\ast 10^{-5}$	SD $\ast 10^{-5}$	Skewness	Kurtosis
Realized Variance						
AIG	17.4	9.16	3.05	11.2	2.47	13.1
AXP	6.12	6.38	1.45	4.71	7.67	91.5
BA	5.46	2.19	1.11	3.25	1.70	7.33
C	16.7	9.62	2.80	10.9	2.56	15.3
Realized Covariance						
AIG-AXP	4.30	1.65	-1.99	3.13	1.00	3.76
AIG-BA	3.32	1.32	-2.74	2.82	1.03	3.93
AIG-C	7.72	4.54	-1.47	5.99	1.85	9.37
AXP-BA	2.41	1.01	-0.59	1.95	1.26	4.37
AXP-C	5.03	2.24	-0.75	3.53	1.61	6.84
BA-C	3.69	1.75	-1.40	3.13	1.66	6.30

Table 5: The Selection of the Order of VAR Model

We fit the VAR model to the vectorized factor covariance matrix $\hat{\Sigma}_f(t)$, which is a vector with 10 entries. Using the package "vars" in R, we select the VAR(1) model as all the model selection criteria, namely Akaike information criterion (AIC), Hannan Quinn (HQ), Schwarz criterion (SC) and final prediction error (FPE), choose the order 1.

Oder	1	2	3	4	5
AIC(n) *10 ²	-1.949	-1.943	-1.937	-1.933	-1.932
HQ(n) *10 ²	-1.939	-1.925	-1.910	-1.898	-1.888
SC(n) *10 ²	-1.924	-1.897	-1.871	-1.846	-1.824
FPE(n) *10 ⁻⁸⁴	0.234	0.416	0.796	1.292	1.751

Table 6: Sparsity of $\hat{\mathbf{A}}_1$

We compute the values $\frac{1}{10^2} \sum_{i,j} |a_{i,j}|^m$ for different $0 \leq m < 1$, where $\hat{\mathbf{A}}_1 = \{a_{i,j}\}_{1 \leq i,j \leq 10}$.

m	0	0.05	0.1	0.15	0.2	0.25	0.3
$\frac{1}{10^2} \sum_{i,j} a_{i,j} ^m$	1	0.8686	0.7586	0.666	0.5877	0.5212	0.4646

Table 7: Forecast errors for CAW and VAR models using different norms

We report the results of the prediction accuracy of the two models using different norms. In addition, the prediction accuracy of the inverse of the matrices is shown as well. Here, FN is for the Frobenius norm and SN for the spectral norm.

CAW					
Order	Number of Parameters	FN $\ast 10^{-4}$	SN $\ast 10^{-4}$	FN for Inverse $\ast 10^5$	SN for Inverse $\ast 10^5$
$p = 0, q = 1$	9	6.10	5.58	7.30	4.31
$p = 1, q = 1$	13	5.27	4.80	7.31	4.31
$p = 1, q = 2$	17	5.28	4.81	7.31	4.31
$p = 2, q = 1$	17	5.28	4.83	7.31	4.31
$p = 2, q = 2$	21	5.28	4.82	7.31	4.31
VAR					
	Number of Parameters	FN $\ast 10^{-4}$	SN $\ast 10^{-4}$	FN for Inverse $\ast 10^5$	SN for Inverse $\ast 10^5$
VAR(1)	110	5.18	4.76	7.31	4.30

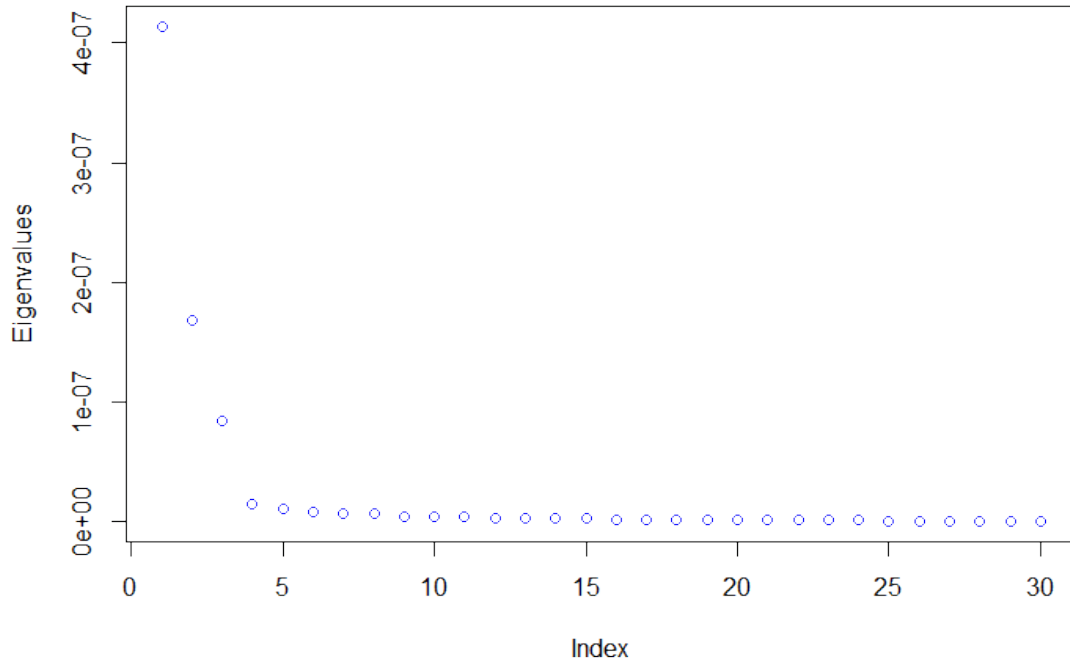


Figure 1: Plots of the eigenvalues of $\tilde{\mathbf{S}}_x$ for the dataset when $k = 98$.

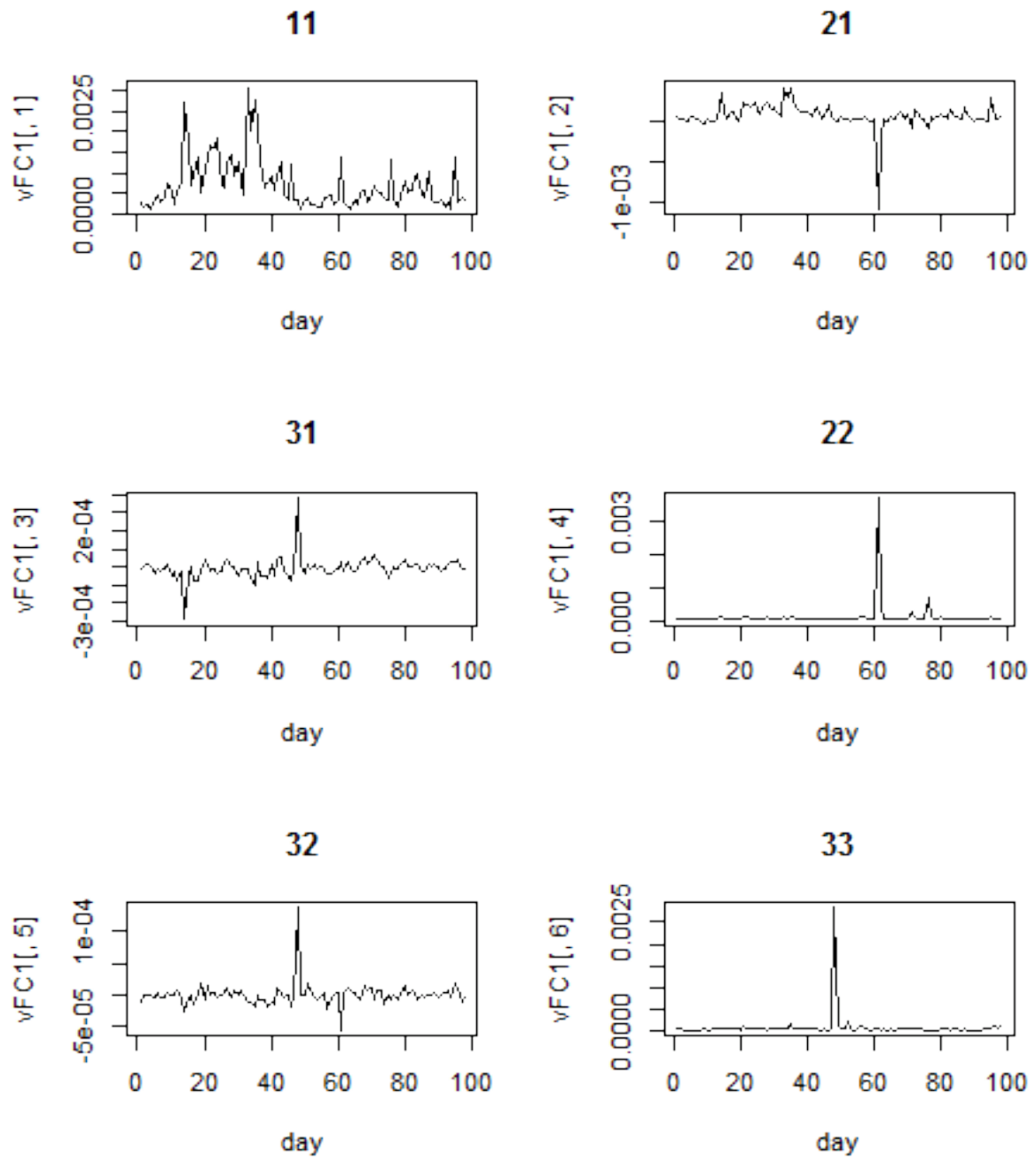


Figure 2: Time series of variances and covariances for factor covariance matrices.

Diagram of fit and residuals for y1

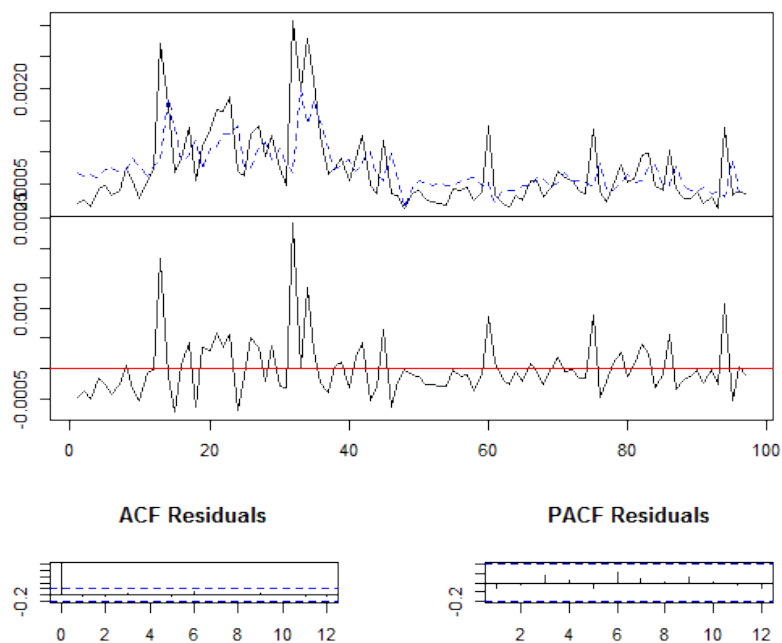


Diagram of fit and residuals for y2

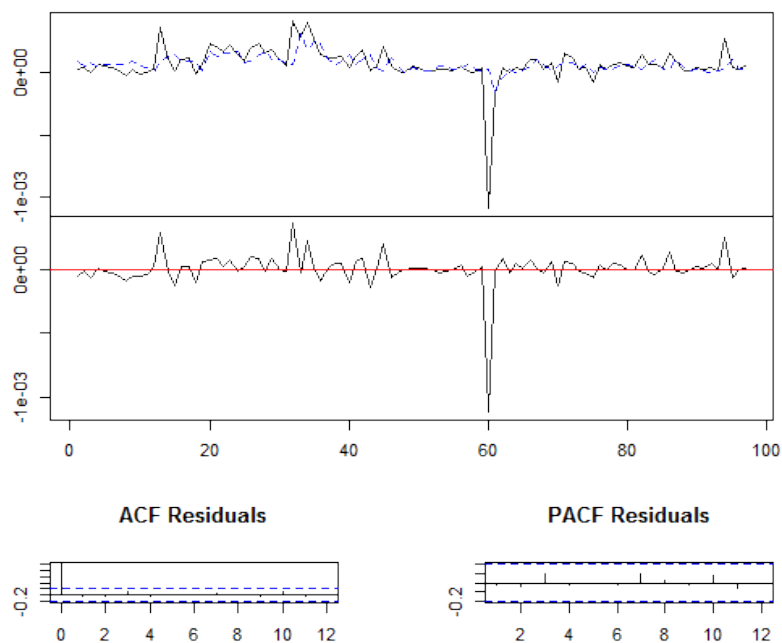


Diagram of fit and residuals for y3

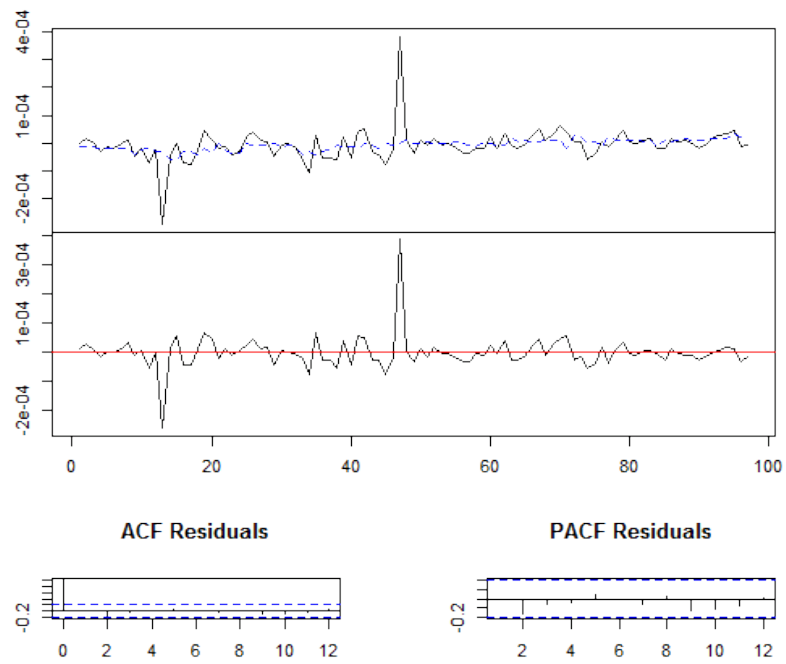
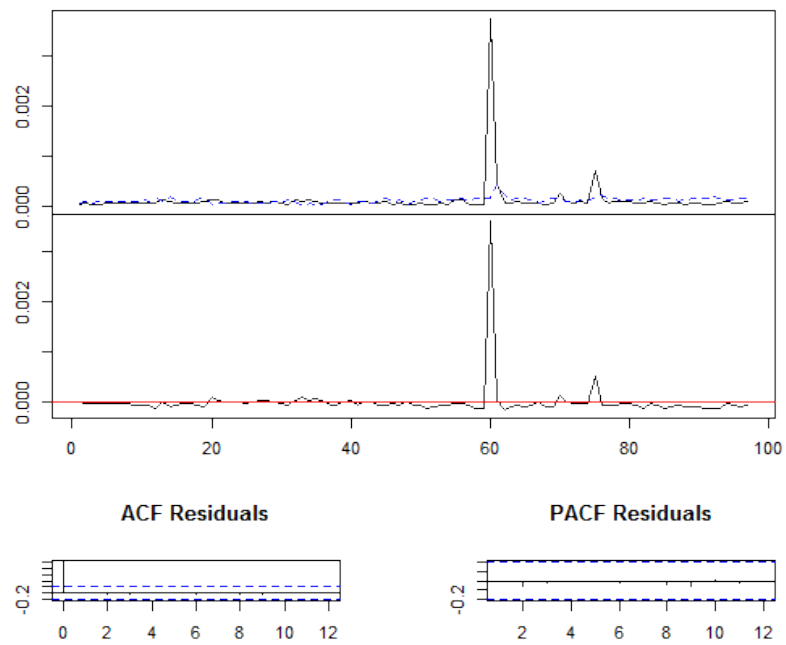


Diagram of fit and residuals for y4



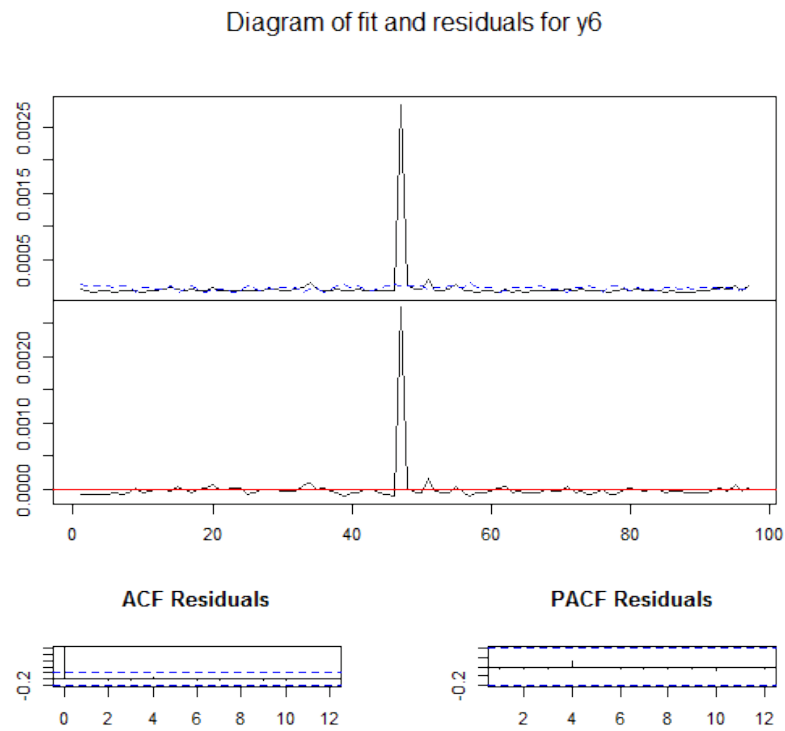
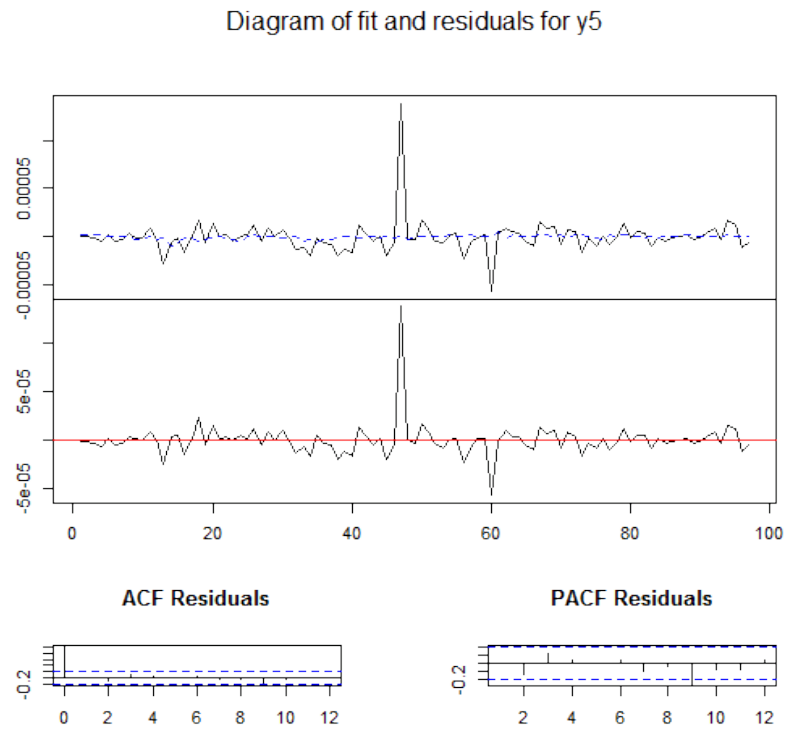
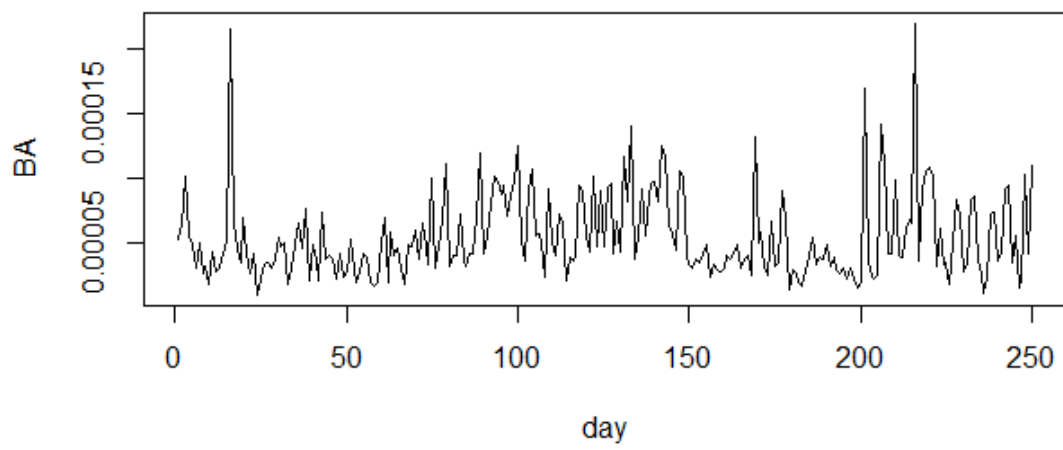
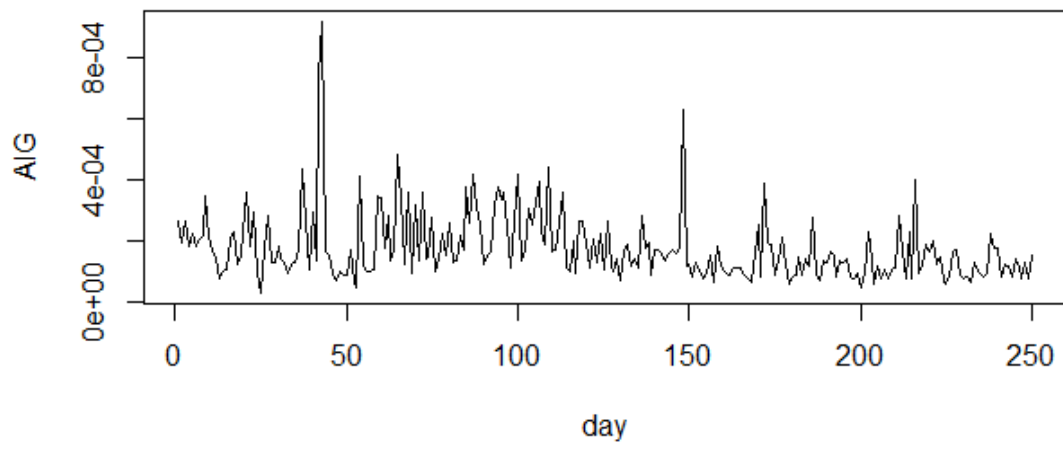


Figure 4: The plot of fitted VAR(1) model, residuals of fitted VAR(1) model, and ACF and PACF of the residuals.



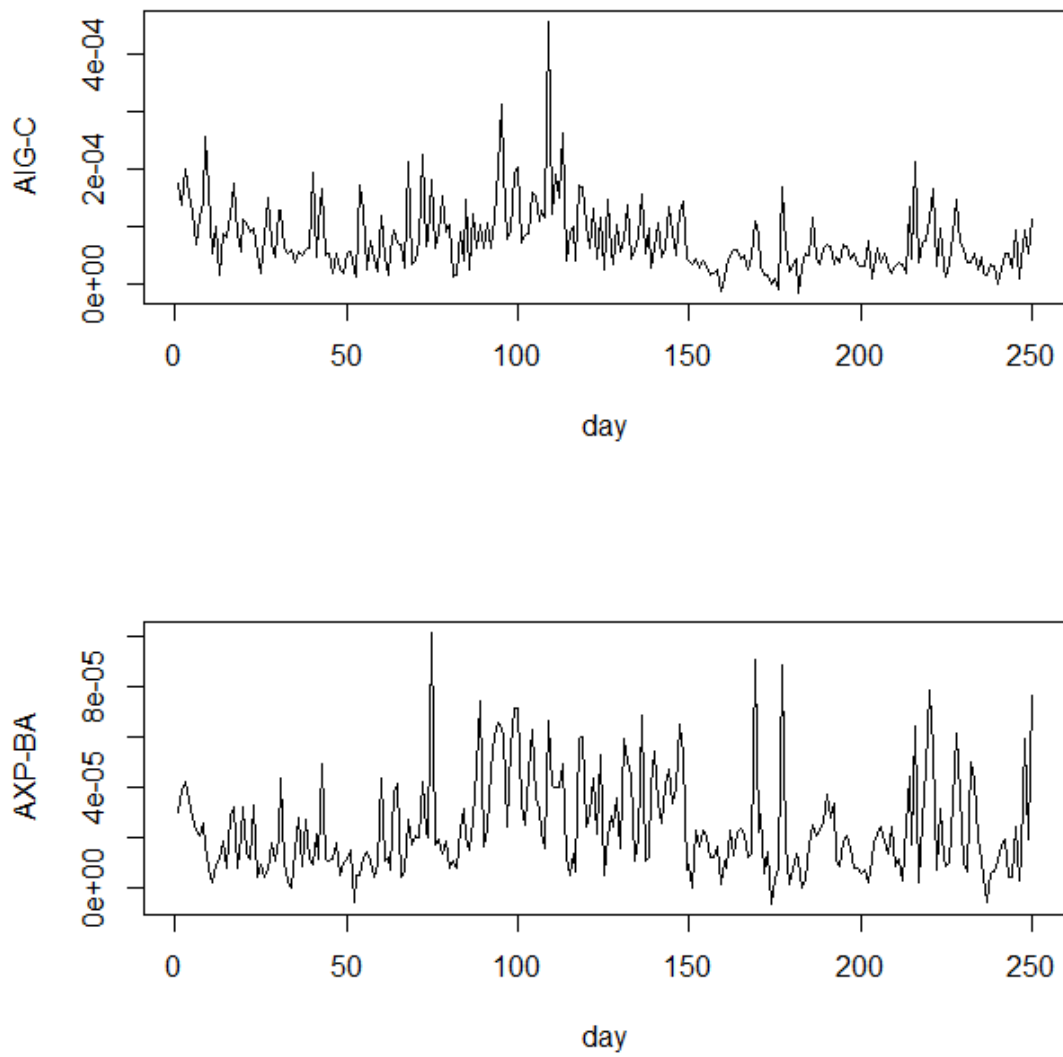


Figure 5: The plot of selected realized variances and covariances.

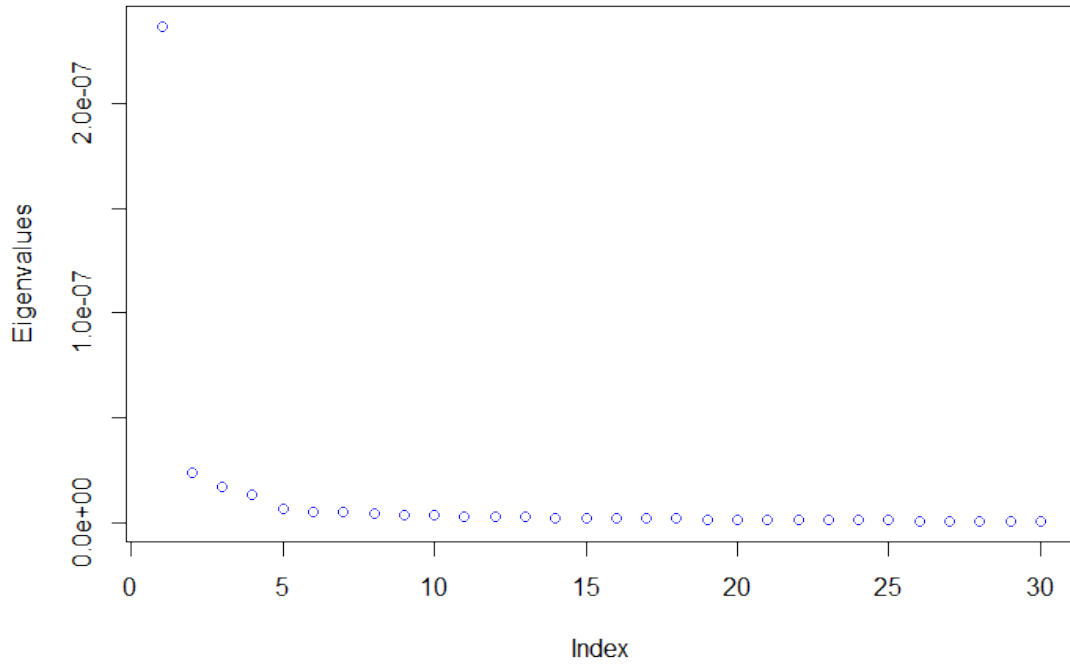


Figure 5: Plots of the eigenvalues of $\hat{\mathbf{S}}_x$ for the dataset when $k = 220$.

Appendix A. Proof of theorems

For convenience, we denote $A(d, T, n)$ and $B(T)$ by A and B in the proof part.

Proof of Theorem 1. Following Theorem 1 in [Tao et al. \(2011\)](#), we can easily show that $\|\tilde{\mathbf{S}}_x - \bar{\mathbf{S}}_x\|_2 = O_p(AB)$.

Then, we claim that

$$\max_{1 \leq j \leq r} |\hat{\lambda}_j - \lambda_j| = O_p(AB) \quad (\text{A.1})$$

$$\max_{1 \leq j \leq r} \|\hat{\mathbf{a}}_j - \mathbf{a}_j\|_2 = O_p(AB) \quad (\text{A.2})$$

Let $P = \tilde{\mathbf{S}}_x - \bar{\mathbf{S}}_x$ with ordered eigenvalues $\rho_1 \geq \dots \geq \rho_d$. Then, we have $\rho_d \leq \hat{\lambda}_j - \lambda_j \leq \rho_1$. As a result,

$$|\hat{\lambda}_j - \lambda_j| \leq \max(|\rho_1|, |\rho_d|) = \|\tilde{\mathbf{S}}_x - \bar{\mathbf{S}}_x\|_2 \quad (\text{A.3})$$

which proves equation (A.1). The equation (A.2) follows from Theorem 1 in [Bickel and Levina \(2008\)](#) and the same argument in the proof of Theorem 5 in the same paper.

For the j -th diagonal entry of $\mathbf{A}'\hat{\mathbf{A}} - \mathbf{I}_r$,

$$\mathbf{a}'_j \hat{\mathbf{a}}_j - 1 = -\|\hat{\mathbf{a}}_j - \mathbf{a}_j\|_2^2 / 2 = O_p(A^2 B^2) \leq O_p(AB) \quad (\text{A.4})$$

as we assume AB goes to zero. For off-diagonal entry (k, j) ($k \neq j$),

$$\begin{aligned} |\mathbf{a}'_k \hat{\mathbf{a}}_j| &= |\mathbf{a}'_k (\hat{\mathbf{a}}_j - \mathbf{a}_j)| \\ &\leq \|\mathbf{a}'_k\|_2 \|\hat{\mathbf{a}}_j - \mathbf{a}_j\|_2 = \|\hat{\mathbf{a}}_j - \mathbf{a}_j\|_2 \\ &= O_p(AB) \end{aligned} \quad (\text{A.5})$$

which proves the first result.

To prove the second result, we separate the left hand side of the equation into three parts:

$$\begin{aligned}
& \hat{\Sigma}_f(t) - \Sigma_f(t) - \mathbf{A}'\Sigma_0\mathbf{A} \\
&= \hat{\mathbf{A}}'[\hat{\Sigma}_x(t) - \Sigma_x(t)]\hat{\mathbf{A}} + \hat{\mathbf{A}}'\Sigma_x(t)\hat{\mathbf{A}} - \Sigma_f(t) - \mathbf{A}'\Sigma_0\mathbf{A} \\
&= \hat{\mathbf{A}}'[\hat{\Sigma}_x(t) - \Sigma_x(t)]\hat{\mathbf{A}} + [(\mathbf{A}'\hat{\mathbf{A}})'\Sigma_f(t)\mathbf{A}'\hat{\mathbf{A}} - \Sigma_f(t)] \\
&\quad + [\hat{\mathbf{A}}'\Sigma_0\hat{\mathbf{A}} - \mathbf{A}'\Sigma_0\mathbf{A}]
\end{aligned} \tag{A.6}$$

For the first term on the right-hand side of equation (A.6), since

$$\|\hat{\mathbf{A}}'\|_2^2, \|\hat{\mathbf{A}}\|_2^2 = 1, \tag{A.7}$$

we have

$$\begin{aligned}
\|\hat{\mathbf{A}}'[\hat{\Sigma}_x(t) - \Sigma_x(t)]\hat{\mathbf{A}}\|_2 &\leq \|\hat{\mathbf{A}}'\|_2 \|\hat{\Sigma}_x(t) - \Sigma_x(t)\|_2 \|\hat{\mathbf{A}}\|_2 \\
&= O_p(A)
\end{aligned} \tag{A.8}$$

For the second term, from Condition (A2), we know that $\|\Sigma_f(t)\|_2 = O_p(B)$, and we have

$$\begin{aligned}
\|\hat{\mathbf{A}} - \mathbf{A}\|_2^2 &= \|(\hat{\mathbf{A}}' - \mathbf{A}')(\hat{\mathbf{A}} - \mathbf{A})\|_2 \\
&\leq 2 \|\mathbf{A}'\hat{\mathbf{A}} - \mathbf{I}_r\|_2 = O_p(AB)
\end{aligned} \tag{A.9}$$

As a result,

$$\begin{aligned}
\|(\mathbf{A}'\hat{\mathbf{A}})'\Sigma_f(t)\mathbf{A}'\hat{\mathbf{A}} - \Sigma_f(t)\|_2 &= \|\hat{\mathbf{A}}'\mathbf{A}\Sigma_f(t)\mathbf{A}'\hat{\mathbf{A}} - \hat{\mathbf{A}}'\hat{\mathbf{A}}\Sigma_f(t)\hat{\mathbf{A}}'\hat{\mathbf{A}}\|_2 \\
&\leq \|\hat{\mathbf{A}}'\|_2 \|\mathbf{A}\Sigma_f(t)\mathbf{A}' - \hat{\mathbf{A}}\Sigma_f(t)\hat{\mathbf{A}}'\|_2 \|\hat{\mathbf{A}}\|_2 \\
&\leq \|(\hat{\mathbf{A}} - \mathbf{A})\Sigma_f(t)\hat{\mathbf{A}}' + \mathbf{A}\Sigma_f(t)(\hat{\mathbf{A}} - \mathbf{A})'\|_2 \\
&\leq \|\hat{\mathbf{A}} - \mathbf{A}\|_2 \|\Sigma_f(t)\|_2 (\|\hat{\mathbf{A}}\|_2 + \|\mathbf{A}\|_2) \\
&= O_p(A^{1/2}B^{3/2})
\end{aligned} \tag{A.10}$$

For the third term, again from Condition (A2), we know that $\|\Sigma_0\|_2$ is bounded, therefore,

$$\begin{aligned}
\|\hat{\mathbf{A}}'\Sigma_0\hat{\mathbf{A}} - \mathbf{A}'\Sigma_0\mathbf{A}\|_2 &= \|(\hat{\mathbf{A}} - \mathbf{A})'\Sigma_0\hat{\mathbf{A}} + \mathbf{A}'\Sigma_0(\hat{\mathbf{A}} - \mathbf{A})\|_2 \\
&\leq \|(\hat{\mathbf{A}} - \mathbf{A})'\|_2 \|\Sigma_0\|_2 \|\hat{\mathbf{A}}\|_2 + \|\mathbf{A}'\|_2 \|\Sigma_0\|_2 \|(\hat{\mathbf{A}} - \mathbf{A})\|_2 \\
&= \|\hat{\mathbf{A}} - \mathbf{A}\|_2 \|\Sigma_0\|_2 (\|\hat{\mathbf{A}}\|_2 + \|\mathbf{A}\|_2) \\
&= O_p(A^{1/2}B^{1/2})
\end{aligned} \tag{A.11}$$

From equations (A.8), (A.10) and (A.11), we conclude that

$$\hat{\Sigma}_f(t) - \Sigma_f(t) - \mathbf{A}'\Sigma_0\mathbf{A} = O_p(A^{1/2}B^{3/2}). \tag{A.12}$$

□

We need the following three lemmas to prove Theorem 2.

Lemma 1. $\partial\mathcal{L}(\theta)/\partial\theta_i$ and $\partial^2\mathcal{L}(\theta)/\partial\theta_i\partial\theta_j$ are uniformly bounded for $\theta \in \Theta$.

Proof. $\mathbf{S}_f(t) \geq CC' > 0$, so that $\mathbf{S}_f(t)^{-1} > 0$. Given $\mathbf{S}_f(0), \mathbf{S}_f(-1), \dots, \mathbf{S}_f(-p+1), \mathbf{S}_f(t)$ is a polynomial of $A_{mn,j}, B_{mn,i}, C_{mn}$'s, for $1 \leq m \leq r$ and $1 \leq n \leq r$, i.e. $\mathbf{S}_f(t) \in C^\infty$. As a result, $\mathbf{S}_f(t)^{-1} \in C^\infty$. We have $\ln(\cdot), \Gamma(\cdot)$ and $\text{tr}(\cdot)$ are all $\in C^\infty$, which indicates $\mathcal{L}(\theta) \in C^\infty$, i.e. $\partial\mathcal{L}(\theta)/\partial\theta_i, \partial^2\mathcal{L}(\theta)/\partial\theta_i\partial\theta_j \in C^\infty$. Since $\theta \in \Theta$ which is compact, we have the conclusion that $\partial\mathcal{L}(\theta)/\partial\theta_i$ and $\partial^2\mathcal{L}(\theta)/\partial\theta_i\partial\theta_j$ are bounded on Θ . □

Lemma 2. If we denote $K(\theta, \tilde{\Sigma}_f(t)) := \hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta) = O_p(A^{1/2}B^{5/2})$, then we have

$$\frac{\partial K(\theta, \tilde{\Sigma}_f(t))}{\partial\theta_i} = O_p(A^{1/2}B^{5/2}). \tag{A.13}$$

Proof. By straightforward calculations, it can be shown that $K(\theta, \tilde{\Sigma}_f(t))$ can be split into:

$$K(\theta, \tilde{\Sigma}_f(t)) = C_1(p, n) + F(\theta) + \sum C_2(p, n)G(\theta) \tag{A.14}$$

where $C_1(p, n) = O_p(A)$ and $C_2(p, n) = O_p(A^{1/2}B^{5/2})$ which are free of θ , while $F(\cdot), G(\cdot) \in C^\infty$, according to Lemma 1. In particular, all the derivatives of F and G are bounded on Θ . As a consequence,

$$\begin{aligned}\frac{\partial K(\theta, \tilde{\mathbf{S}}_f(t))}{\partial \theta_i} &= \frac{\partial F(\theta)}{\partial \theta_i} + \sum C_2(p, n) \frac{\partial G(\theta)}{\partial \theta_i} \\ &= O_p(A^{1/2}B^{5/2}).\end{aligned}\tag{A.15}$$

□

Lemma 3. *The log-likelihood function for the CAW model given observed data $\hat{\mathbf{S}}_f(t)$ and true data $\tilde{\mathbf{S}}_f(t)$ are*

$$\begin{aligned}\hat{\mathcal{L}}(\theta) &= \frac{1}{T} \sum_{t=1}^T \left\{ -\frac{\nu r}{2} \ln(2) - \frac{r(r-1)}{4} \ln(\pi) - \sum_{i=1}^r \ln \Gamma\left(\frac{\nu+1-i}{2}\right) - \frac{\nu}{2} \ln \left| \frac{\hat{\mathbf{S}}_f(t)}{\nu} \right| \right. \\ &\quad \left. + \left(\frac{\nu-r-1}{2}\right) \ln |\hat{\mathbf{S}}_f(t)| - \frac{1}{2} \text{tr}(\nu \hat{\mathbf{S}}_f(t)^{-1} \hat{\mathbf{S}}_f(t)) \right\}\end{aligned}\tag{A.16}$$

and

$$\begin{aligned}\mathcal{L}(\theta) &= \frac{1}{T} \sum_{t=1}^T \left\{ -\frac{\nu r}{2} \ln(2) - \frac{r(r-1)}{4} \ln(\pi) - \sum_{i=1}^r \ln \Gamma\left(\frac{\nu+1-i}{2}\right) - \frac{\nu}{2} \ln \left| \frac{\mathbf{S}_f(t)}{\nu} \right| \right. \\ &\quad \left. + \left(\frac{\nu-r-1}{2}\right) \ln |\tilde{\mathbf{S}}_f(t)| - \frac{1}{2} \text{tr}(\nu \mathbf{S}_f(t)^{-1} \tilde{\mathbf{S}}_f(t)) \right\},\end{aligned}\tag{A.17}$$

respectively. Then we have

$$\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta) = O_p(A^{1/2}B^{5/2})\tag{A.18}$$

Proof. By simple algebraic manipulations, we have

$$\begin{aligned}\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta) &= \frac{1}{T} \sum_{t=1}^T \left\{ -\frac{\nu}{2} (\ln \left| \frac{\hat{\mathbf{S}}_f(t)}{\mathbf{S}_f(t)} \right|) + \left(\frac{\nu-r-1}{2}\right) \ln \left| \frac{\hat{\mathbf{S}}_f(t)}{\tilde{\mathbf{S}}_f(t)} \right| \right. \\ &\quad \left. - \frac{1}{2} \text{tr}(\nu (\hat{\mathbf{S}}_f(t)^{-1} \hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t)^{-1} \tilde{\mathbf{S}}_f(t))) \right\}\end{aligned}\tag{A.19}$$

Since $\mathbf{S}_f(t)$ is the conditional mean of $\tilde{\mathbf{S}}_f(t)$, it can be proved that $\max_t \|\mathbf{S}_f(t)\|_2 = O_p(B)$ and $\hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t) = O_p(A^{1/2}B^{3/2})$ as $\tilde{\mathbf{S}}_f(t)$ and $\hat{\mathbf{S}}_f(t) - \tilde{\mathbf{S}}_f(t)$. The basic idea of the proof is the fol-

lowing. We can treat Equation (10) as a linear map from $(\tilde{\mathbf{S}}_f(t))$ to $(\mathbf{S}_f(t))$: $(\mathbf{S}_f(t)) = \mathcal{L} \cdot (\tilde{\mathbf{S}}_f(t))$ where \mathcal{L} is a linear operator with bounded norm. It follows that $\forall t, \|\mathbf{S}_f(t)\|_2 \leq \alpha \sup_s \|\tilde{\mathbf{S}}_f(t)\|_2$ for some constant α , which gives the required bound. A similar argument can be applied to $\hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t)$.

For the first term of the additional part, since $\mathbf{S}_f(t) = O_p(B)$, we have

$$\left\| \frac{\hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t)}{\mathbf{S}_f(t)} \right\|_2 = O_p(A^{1/2}B^{1/2}) \quad (\text{A.20})$$

Thus,

$$\ln \left| \frac{\hat{\mathbf{S}}_f(t)}{\mathbf{S}_f(t)} \right| = \left(\left| \frac{\hat{\mathbf{S}}_f(t)}{\mathbf{S}_f(t)} \right| - 1 \right) - o_p \left(\left| \frac{\hat{\mathbf{S}}_f(t)}{\mathbf{S}_f(t)} \right| - 1 \right) = O_p(A^{1/2}B^{1/2}) \quad (\text{A.21})$$

For the second term, it is easy to prove that $\tilde{\mathbf{S}}_f(t) = O_p(B)$, so that we have

$$\left\| \frac{\hat{\mathbf{S}}_f(t) - \tilde{\mathbf{S}}_f(t)}{\tilde{\mathbf{S}}_f(t)} \right\|_2 = O_p(A^{1/2}B^{1/2}) \quad (\text{A.22})$$

Thus,

$$\ln \left| \frac{\hat{\mathbf{S}}_f(t)}{\tilde{\mathbf{S}}_f(t)} \right| = \left(\left| \frac{\hat{\mathbf{S}}_f(t)}{\tilde{\mathbf{S}}_f(t)} \right| - 1 \right) - o_p \left(\left| \frac{\hat{\mathbf{S}}_f(t)}{\tilde{\mathbf{S}}_f(t)} \right| - 1 \right) = O_p(A^{1/2}B^{1/2}) \quad (\text{A.23})$$

For the third term, first note that as $\mathbf{S}_f(t) \geq CC'$, there exists a constant $w > 0$ such that the minimum eigenvalue of $\mathbf{S}_f(t)$ is larger than or equal to w . Consequently, $\|\mathbf{S}_f(t)^{-1}\|_2 \leq \frac{1}{w}$ which is bounded. As a result

$$\begin{aligned} & \|\hat{\mathbf{S}}_f(t)^{-1}\hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t)^{-1}\tilde{\mathbf{S}}_f(t)\|_2 \\ & \leq \|\hat{\mathbf{S}}_f(t)^{-1}(\hat{\mathbf{S}}_f(t) - \tilde{\mathbf{S}}_f(t))\|_2 + \|(\hat{\mathbf{S}}_f(t)^{-1} - \mathbf{S}_f(t)^{-1})\tilde{\mathbf{S}}_f(t)\|_2 \\ & \leq \|\hat{\mathbf{S}}_f(t)^{-1}\|_2 \|\hat{\mathbf{S}}_f(t) - \tilde{\mathbf{S}}_f(t)\|_2 + \|\hat{\mathbf{S}}_f(t)^{-1}\|_2 \|\hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t)\|_2 \|\mathbf{S}_f(t)^{-1}\|_2 \|\tilde{\mathbf{S}}_f(t)\|_2 \\ & = \frac{1}{w} O_p(A^{1/2}B^{3/2}) + \frac{1}{w^2} O_p(A^{1/2}B^{3/2}) O_p(B) \\ & = O_p(A^{1/2}B^{5/2}) \end{aligned} \quad (\text{A.24})$$

□

Proof of Theorem 2. From Lemma 3, we have $\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta) = K(\theta, \tilde{\mathbf{S}}_f(t)) = O_p(A^{1/2}B^{5/2})$. By

taking derivatives of both sides with respect to any parameters from θ , we have

$$\frac{\partial \hat{\mathcal{L}}(\theta)}{\partial \theta_i} = \frac{\partial \mathcal{L}(\theta)}{\partial \theta_i} + \frac{\partial K(\theta, \tilde{\Sigma}_f(t))}{\partial \theta_i} \quad (\text{A.25})$$

By Lemma 2, we have

$$\begin{aligned} 0 = \frac{\partial \hat{\mathcal{L}}(\hat{\theta})}{\partial \theta_i} &= \frac{\partial \mathcal{L}(\hat{\theta})}{\partial \theta_i} + O_p(A^{1/2} B^{5/2}) \\ &= \frac{\partial \mathcal{L}(\tilde{\theta})}{\partial \theta_i} + \sum_j \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j}(\tilde{\theta})(\hat{\theta} - \tilde{\theta}) + o_p(\hat{\theta} - \tilde{\theta}) + O_p(A^{1/2} B^{5/2}) \\ &= \sum_j \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j}(\tilde{\theta})(\hat{\theta} - \tilde{\theta}) + o_p(\hat{\theta} - \tilde{\theta}) + O_p(A^{1/2} B^{5/2}) \end{aligned} \quad (\text{A.26})$$

From Lemma 1, we know that $\partial^2 \mathcal{L}(\tilde{\theta})/\partial \theta_i \partial \theta_j$ are bounded for any $\tilde{\theta}$. In addition, from Conditions (A4) and (A6), there exists some constant ℓ such that in probability,

$$|\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j}(\tilde{\theta})| > \ell > 0 \quad (\text{A.27})$$

uniformly. As a result, we have

$$\hat{\theta} - \tilde{\theta} = O_p(A^{1/2} B^{5/2}) \quad (\text{A.28})$$

□

Figure Legends:

Figure 2:

The subfigure with title "11" shows the time series of the variance for the first factor, the subfigure with title "21" shows that of the covariance between the first and second factors, and so on.

Figure 3:

The first subplot is for the first entry of the 6-dimensional vector, and so on. For each of the 6 time series, we show the in-sample fit in the upper panel, where the solid line stands for the real data and the dashed line is the fitted series. The residuals are shown in the middle panel while the ACF and PACF of residuals are displayed in the lower panel.

Figure 4:

Two plots for realized variances and two plots for realized covariances are shown in Figure 5. AIG and BA are chosen for the realized variances while the realized covariances between AIG and C, and between AXP and BA are shown in the following figures.

References

- T.G. Andersen, T. Bollerslev, F.X. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71:579–625, 2003.
- M. Asai and M.J. McAleer. Forecasting co-volatilities via factor models with asymmetry and long memory in realized covariance. (No. TI 2014-037/III). *Tinbergen Institute Discussion Paper Series.*, 2014.
- K. Bannouh, M. P. E. Martens, R. C. Oomen, and D. J. Van Dijk. Realized mixed-frequency factor models for vast dimensional covariance estimation. (No. ERS-2012-017-F&A). *ERIM Report Series Research in Management.*, 2012.
- O. E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72:885–925, 2004.
- O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard. Realized kernels in practice: Trades and quotes. *Journal of Econometrics*, 12(3):C1–C32, 2009.
- O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162(2):149–169, 2011.
- P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.
- T. Bollerslev, R.F. Engle, and J. Wooldridge. A capital asset pricing model with time varying covariances. *Journal of Political Economy*, 95:116–131, 1988.
- C. T. Brownlees and G. M. Gallo. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis*, 51(4):2232–2245, 2006.
- K. Christensen, S. Kinnebrock, and M. Podolskij. Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics*, 159(1):116–133, 2010.

- R.F. Engle and K.F. Kroner. Multivariate simultaneous generalized arch. *Econometric Theory*, 11:122–150, 1995.
- J. Fan and Y. Wang. Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, 102(480):1349–1362, 2007.
- V. Golosnoy, B. Gribisch, and R. Liesenfeld. The conditional autoregressive wishart model for multivariate stock market volatility. *Journal of Econometrics*, 167(1):211–223, 2012.
- C. Gourieoux, J. Jasiak, and R. Sufana. The wishart autoregressive process of multivariate stochastic volatility. *Journal of Econometrics*, 150(2):167–181, 2009.
- T. Hayashi and N. Yoshida. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2):359–379, 2005.
- X. Jin and J. M. Maheu. Modelling realized covariances. (*No. tecipa-382*), 2009.
- I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.
- H.M. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.
- T. H. McCurdy and T. Stengos. A comparison of risk-premium forecasts implied by parametric versus nonparametric conditional mean estimators. *Journal of Econometrics*, 52(1):225–244, 1992.
- F.C. Ng, W.K. Li, and P.L.H. Yu. The generalized conditional autoregressive wishart model for multivariate stochastic volatility. *Working Paper*, 2014.
- M. Tao, Y. Wang, Q. Yao, and J. Zou. Large volatility matrix inference via combining low-frequency and high-frequency approaches. *Journal of the American Statistical Association*, 106(495), 2011.
- M. Tao, Y. Wang, and X. Chen. Fast convergence rates in estimating large volatility matrices using high-frequency financial data. *Econometric Theory*, 29(04):838–856, 2013.

- Y. Wang and J. Zou. Vast volatility matrix estimation for high-frequency financial data. *The Annals of Statistics*, pages 943–978, 2010.
- L. Zhang. Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 12(6):1019–1043, 2006.
- L. Zhang. Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47, 2011.